DHCANVAS: a Web Application for Exploration and Annotation of Historical Documents

Orlin Topalov, DHLAB, orlin.topalov@epfl.ch
Maud Ehrmann, DHLAB, maud.ehrmann@epfl.ch
Frédéric Kaplan, DHLAB, frederic.kaplan@epfl.ch

A web application for the visualization, the transcription and the annotation of historical documents.

Context
- massive digitization of archival documents
- need to extract information out of this material

Objective
allow users to collaboratively work on documents and to produce information which can be stored, processed, retrieved and exchanged.

Passive visualization
- one page display
- zoom
- rotation
- filmstrip
- navigator

Upcoming
- sequence mode
- collection mode
- book mode

Active visualization
- create segments
- resize
- group/ungroup
- delete

Upcoming
- segment linking
- merge/split
- drag-select

Visualizaion capacities
- IIIF image and presentation API
- Loris image server
- OpenSeadragon viewer

Annotation capacities
The user reads a page, selects an image segment and creates annotations about the segment.

Annotation model
Annotation objects encode information which belongs to different levels:
- Resource level — what relates to a digital resource/canvas, e.g. an image segment representing the name of a person.
- Document level — what relates to textual content, e.g. an annotation representing the mention of a person.
- Entity level — what relates to concepts/entities of the world, e.g. an annotation representing the entity a mention refers to.

Annotation workflow
1. the user selects a segment and transcribes it. The transcription annotation consists of:
- literal transcription (mandatory)
- language used (mandatory)
- transcription certainty (opt.)
- standard form (opt.)

2. the user describes what the text is about. The mention annotation consists of:
- a semantic category: broad category of things to which the mention belongs to (mandatory)
- a semantic tag: sub-category which specifies the mention locally
- a series of attributes

3. the user identifies to which real-world entity the mention refers to. The entity annotation consists of:
- a name, which should be unique
- possibly a relationship towards another entity

Representation & Implementation
- Shared Canvas Data Model & Open Annotation Data Model
- PostgreSQL and JSON schemas

Main functionalities
- visualization
- transcription
- annotation
- search

Key Features
- preserve the link between images and annotations
- allow concurrent annotations on the same object
- fully compliant with standards (IIIF, SC, OA)

Search Capacities
- segment-based
- faceted search
- weighted full-text search

Future Works
- open source release on GitHub
- export module (IIIF, OA, SKOS and RDF compatible)
- complementary service integration via REST APIs e.g.: line detection, segmentation, handwritten text recognition
- metadata and vocabulary edition capacities
- user role and permission management

EPFL IC Research Day, June 30th, 2016